J-0152
C.2

Health Hazards in Mining

The Files and Facilities

Martha E. Smith
Head, Occupational and Environmental Research Unit
Vital Statistics and Disease Registries Section
R.H. Coats Building
Tunney's Pasture
Ottawa, Ontario
K1A 0Z5
(613) 995-9593

A paper presented at the international conference "Radiation Hazards in Mining: Control, Measurement, and Medical Aspects" held October 4-9, 1981, Colorado School of Mines, Golden, Colorado. The conference proceedings are to be published. To purchase copies of the proceedings write to:

Ms. Marianne Snedeker
Manager of Publications
Society of Mining Engineers of AIME
Caller No. D
Littleton, CD 80127
Phone: (303) 973-9550

HEALTH HAZARDS IN MINING
THE FILES AND FACILITIES

Martha E. Smith

Vital Statistics and Disease Registries
Statistics Canada
Ottawa, Ontario

INTRODUCTION

Radiation is one of the many agents occurring in our environment that is capable of causing cancers. We are all unavoidably exposed to natural background radiation.

If we wish to exploit the beneficial uses of radiation in medicine, industry, and for the production of electricity, some further exposures to radiation are almost inevitable. One needs to ensure therefore, that the standards of radiation protection are safely derived and one would want to understand the mechanism by which radiation might cause cancer and genetic defects.

In order to study the effects of radiation one can: (1) analyze human data, (2) perform animal experiments, and (3) do molecular level experiments. Indeed all three kinds of information are needed. This paper will concentrate on the steps required to obtain some of the necessary human data regarding health effects, where the concern is about the impact and consequences of long-term low-level exposures to an agent. Such investigations require some knowledge of work histories, dose histories, health "outcomes", and the personal identification of the individual involved.

Three inter-related computer systems have been developed at Statistics Canada which have been designed to permit optimal use of a number of different records for our entire country for such health related research. The development of the Canadian Mortality Data Base, the initiation of the National Cancer Incidence Reporting System, and the development of new computer linkage techniques have helped reduce the cost and increase the scale and efficiency of automated follow-up to produce statistics of sickness or death associated with radiation and other carcinogenic agents in mines. These computer systems have already been implemented, and references will be made to studies currently being conducted using these files and facilities (e.g. a study of all Ontario miners, plus various Canadian uranium, fluorospar, salt and nickel miners). We will also look at the kinds of data that need to be collected now, to improve such studies in the future.

DELAYED RISKS - THE STUDY SIZE AND COST

Delayed effects on human health, as for example industrially caused cancer, can in general only be detected and measured by following-up the individuals to see what eventually becomes of them.

What is not generally recognized is that the relatively low levels of individual risk, about which the public is often concerned, usually requires for their detection that very large numbers of "exposed" and "control" individuals (e.g. 10,000 to 100,000 or more) be followed over a period of two or three decades to determine when they die, what they die of, and whether they contracted cancer or some other disease of special concern. Thus, it is frequently exceedingly difficult to make such investigations cost-effective so that they will be undertaken at all, and as a result very real risks to health can remain undetected or unquantified.

THE MANUAL PROCEDURES

Follow-up of individuals by epidemiologists has until recently been a largely manual and clerical operation. It has used a diversity of source record files; local, regional, and national. Often the tracing of people has involved letters sent through the mails and visits to institutions, physicians, municipal offices, and former neighbours. Only thus could one find out whether the individuals were dead or alive. Such studies were necessarily small, or else very expensive.

Death registrations have, by tradition, provided a valuable tool for the identification of harmful influences in the environment. However, with the increasing mobility in the population, death may often occur far away from the place of exposure to such an influence. Thus, no longer will simple manual searches in a single registry office suffice to inform the investigator concerning the deaths that may have occurred in a study population, especially when that population is large.

In the past, manual follow-up to locate the relevant death registration normally required some prior knowledge of the province and year of death in question, so that the alphabetic indexes could be used to direct researchers to the appropriate bound volumes of registration forms. To search manually in this fashion for any large number of death registrations, without knowing in which year or province the deaths had taken place, or even whether they had yet occurred, would be impractical, since each year in Canada there are about 170,000 deaths.

DEATH AND CANCER AS SPECIAL ENDPOINTS

Much of recent effort has been focused on the organization of the "endpoint" files required to do long-term follow-up studies on a national scale, because that is a function which other institutions are unable to perform due to the confidentiality laws governing the use of such information. Outside organizations generally come to us with detailed "starting" point records which relate to some specific group requiring study. We carry out these epidemiological searches on a cost-recovery basis. The analytical interpretation of results is normally

done by an outside organization, such as the National Cancer Institute of Canada, federal and provincial organizations, and universities.

The Mortality Data Base is maintained in three forms:

The first consists of machine-readable tapes with over 4 million events for the period 1950-77 with coded cause of death. The records are in a standard format, consolidated over all Canadian provinces, and over a span of years to facilitate searching by computer. The file is growing at a rate of about 170,000 events per year and is backed up by source documents on microfilm. Work is in progress to standardize data for the earlier period 1926-49.

The second is compact microfiche, alphabetic listings dating back to 1926. This is used for small research studies (the coded cause of death is missing prior to 1950). Two additional sets have been created from 1950 onward, in alternative sort sequences. One of these is sorted by death year, province, and registration number and the other is in order of death date and surname spelling. The latter file is particularly useful where date of death is known.

The third consists of an historical summary file which has been created to serve special requests for historical tabulations. A typical use of such a file is the production of mortality atlas data to illustrate the spatial variation of mortality rates in Canada in order to facilitate the identification of high risk regions, and the detection of any noteworthy patterns of disease distribution (Minister of Supply and Services Canada 1980 and 1980a). This file is also useful for deriving appropriate denominator figures broken down in a fashion which represents a refinement of the regular published statistics. Control group data may, for example, be derived broken down by age, sex, and small geographic area.

The National Cancer Incidence Reporting System, in operation since 1969, is based on reports from provincial cancer registries, and covers all of Canada except Ontario. It is anticipated that Ontario will become available shortly. Each province in Canada has a central cancer registry which tries to record each newly diagnosed primary cancer. Most provincial registries are also responsible for recording treatment and other follow-up information. About 90% of the data arrives at Statistics Canada on tape and the remaining 10% comes in on simple "report of malignant neoplasm" cards. There are approximately 50,000 notifications each year. The total file contains about 500,000 cancer registrations, and although there have been slight changes over the years in the detail of items or in their definitions, the file is in a standard format for the entire period.

To date, because this is a relatively new system, less experience has been gained in utilizing the cancer incidence files for medical follow-up studies. Some studies are now in the planning stage. This file will be particularly valuable for cancers where the survival rate is high (e.g. breast cancer, bladder cancer, and cancer of the uterus) or where the cancer is rare.

GENERALIZED ITERATIVE RECORD LINKAGE SYSTEM

A generalized record linkage system, which applies probabilistic matching techniques and uses new data base technology and utilities, has been developed and successfully used for a wide variety of research applications. This system has been described in detail elsewhere (Howe and Lindsay 1981; Smith and Silins 1981).

Some of the major features of the system are that it operates in either batch or on-line mode, is modular in development, has been designed to be simple to use, and utilizes weights to produce a quantitative measure of the total probability that two records being compared do or do not relate to the same entity. The particular rules used in the linkage are tailored to the files coming to us from a wide variety of research areas. A flexible command language has been developed to control the individual steps in the linkage process.

In developing a computer strategy for linking records, it was found that the rules of subjective judgement, which a human clerk uses when matching records, are fortunately simpler to simulate by computer than one might originally have supposed.

To get around spelling variations in surname, a phonetic New York State Identification and Intelligence System (NYSIIS) code is used. To aid in deciding whether pairs of records do or do not relate to the same individual, we use the fact that rare names and rare birthplaces carry more discriminating power when they agree than do their commoner counterparts. For example, if we had two records with "J. Smith" on them, subjectively, we would not be sure that they relate to the same individual, but if the name were "Zacharias Orvil Quigley" we would have greater assurance the pair belongs to the same individual. Numeric weights are assigned for agreements and disagreements. Logarithms to the base two are used as in information theory, the weights are additive, and a total weight is achieved at the end indicating whether the pairs are definitely linked, possibly linked, or rejected.

THE PRODUCTS

Most of the studies undertaken to-date relate to groups of people numbering in the tens of thousands to hundreds of thousands. The industrial groups include uranium miners, hard rock and salt miners, nickel workers, uranium refinery workers, radiation workers, plus some who are exposed to asbestos, fibreglass, vinyl chloride, and formaldehyde vapour. Table 1 indicates the wide range of topics of interest and also the large scale of the studies being carried out. An expanded list of references for these studies is given elsewhere (Smith 1979; Smith et al. 1980; Smith and Newcombe 1981; Smith 1981a). Descriptions of the Ontario miners study, Eldorado uranium workers, Newfoundland fluorspar miners, and radiation workers appear elsewhere in the proceedings.

THE FUTURE - IMPROVING THE STARTING POINT DATA SOURCES

The chief limiting factor, where information is sought concerning the health impacts of possible harmful agents in the workplace and elsewhere, and of various social and economic characteristics, has to do with a shortage of "starting" point records which would identify large enough numbers of "exposed" individuals.

Table 1 - The Size and Type of Long-Term Follow-up Studies

Study	Number of Individuals
A Occupational Groups	
1. Ontario uranium miners	16,000
2. Canadian labour force - 10% sample	700,000
3. INCO nickel workers	62,000
4. Falconbridge nickel workers	12,000
5. Eldorado uranium workers	16,000
6. Ontario nominal roll	57,000
- uranium miners	
- non-uranium miners	
- salt miners	
- asbestos workers	
- insulation workers (fibreglass)	
- nickel sinter workers	
- nickel carbonyl workers	
- polyvinyl chloride workers	
- vinyl chloride workers	
7. Newfoundland fluorspar miners	2,000
8. Railway workers	18,000
9. Ontario morticians	1,500
10. Asbestos workers	2,000
B Medical Diagnosis and Treatment	
1. Isoniazid and cancer in tuberculosis patients	64,000
2. Fluoroscopy and cancer in tuberculosis patients	100,000
C Reproductive Problems	
1. Infant death - birth linkage (1971 births)	6,000
2. Infant death - birth linkage (1978 births)	4,200
3. Breast cancer and age at first birth	300,000
D Lifestyle and Other Environmental Factors	
1. Nutrition Canada Survey participants	20,000
E Cancer Incidence, Prevalence, Evaluation of Care, and Survival	
1. Ontario Cancer Registry Reporting System	125,000
2. Alberta Cancer Registry Death Clearance	175,000

Table 2 - A List of Items to be Included in Any "Employee Health Identity Questionnaire"*

1. Surname
2. Alternate surname (if any)
3. First given Name
4. Second and other given names
5. Usual name (or nickname)
6. Sex
7. Birth date (year, month, day)
8. Birth province or country
9. Birth city or place
10. Father's surname
11. Father's first given name
12. Father's second given name
13. Father's birth province or country
14. Mother's maiden surname
15. Mother's first given name
16. Mother's second given name
17. Mother's birth province or country
18. Marital status
19. Spouse's birth surname
20. Spouse's first given name
21. Spouse's second given name
22. Spouse's birth province or country
23. Social Insurance Number
24. Health Insurance Number
25. Pension plan number
26. Signature
27. Date (year, month, day)
28. Complete address including postal code

* This form is designed to be filled out when the employee is hired. Other information relating to work histories and updates to the questionnaire should be added later. Termination dates should be added. A "last known alive date" is of value in order to reduce the amount of searching required in the death file. The complete address should be updated wherever possible.

Although the absence of name rosters of past employees in industry makes it difficult to obtain early evidence of delayed effects amongst employees, there is much that can be done prospectively. A number of possible approaches, using various kinds of starting point records, will be mentioned.

(a) Name rosters of present and future employees, for health follow-up, are being developed by some employers in some industries, especially in connection with uranium mining and refining, and radiation workers. An important consideration in the design of these is the inclusion of an abundance of personal identifying information to take care of problems arising from errors and discrepancies in reporting and recording of such things as names and birth dates (see Table 2).

(b) What is important as a refinement in all comparisons involving supposed at-risk groups, especially where the endpoint (e.g. death) is remote in time, is that confounding factors be controlled in as much as the available information concerning the characteristics of the individuals permit. Demographic and socio-economic characteristics are of particular importance in this connection, and no single centralised kind of record documents these better than the census enumeration forms. The sensitivity of detection of small mortality differences is very much dependent on the size of the study population. But the personal identifying information is largely lacking from the census magnetic tapes, and keying such additional information to tape as a special operation, for the whole of the country, would be excessively expensive. Furthermore, there would have to be some special and well-recognized need for such a special operation to be acceptable. Given such a need, however, it would be feasible to do the keying to tape only for those communities in which supposedly hazardous substances existed. An example would be the former residents of homes recently found to have high radon levels, as compared with a suitable control population. Specific topics of interest might also be determined by examining the mortality atlas for high risk geographic areas. Census data could also be used to confirm the

completeness of a cohort population in a given geographic area, as compiled from various other data sources.

(c) Workmen's Compensation records of employees have already been used in the province of Ontario for follow-up studies of uranium and other hard-rock miners, and this practice could well be extended in the future to other provinces.

(d) Follow-up studies of many sorts of population sub-groups are becoming increasingly popular. For example, a roster exists in the form of the National Dose Registry with information regarding radiation workers (Ashmore 1979; Newcombe 1980; Dreyer et al. 1980). This Registry is now in the process of being adapted for epidemiological follow-up. Moreover, there is new provincial legislation in Ontario which specifically requires the gathering of information and its retention for specific "high" risk groups.

(e) Other survey data such as the Canada Health Survey, the Labour Force survey, medical questionnaires, and smoking data may assist either in providing the basis for new studies, or in establishing better comparison groups so as to control confounding factors in the analysis of the data.

(f) Miscellaneous sources that might sometimes be of use relating to particular regions, include: (i) municipal tax rolls and utility billing lists; (ii) commercial city directories; (iii) Post Office household directories for regions not covered by city directories; (iv) voter's lists for federal and provincial elections; (v) telephone directories; and (vi) driver's licence files.

THE FUTURE - IMPROVING THE ENDPOINT DATA SOURCES

Endpoints that are of major epidemiologic interest include genetic defects, reproductive problems, cancer, and mortality.

One useful indicator of the effectiveness of a new epidemiological program or facility is the demand from, and funding made available by, outside organizations for cost-recovery work. As researchers have heard about the new files and facilities being developed in Canada, and realize their potential, there has been an increasing demand for such studies far beyond our original expectation. Institutions are now planning, budgeting, and requesting that studies be carried out as far ahead as 1984.

Experience in using the Mortality Data Base file has indicated that an extended machine-readable format would be desirable in the near future. For example, the occupation and industry information as reported on the source death registrations are not currently being coded. It was recommended both at our recent Vital Statistics Council meeting and the National Cancer Registry meeting, that steps be taken to have this information coded at a national level. Further particulars recorded on the source document such as city or place of birth, complete parental given names, full spouse names, residence, and postal codes would be helpful to have in machine-readable form for the purpose of increasing the accuracy of the death searches. Birth date information in particular should be improved, and the Social Insurance Number should be incorporated if possible. Current facilities could be extended back to 1926 to include cause of death, and information on the war dead should be put into machine-readable form.

The National Cancer Incidence Reporting System requires additional identifying items to facilitate linkage. A pilot study using the Alberta cancer registry file is currently being carried out to evaluate the feasibility of linkage to the mortality file to obtain survival data. It can be anticipated that there will be problems when surname changes have occurred. Some provincial registries are currently redesigning their data collection procedures, so this problem may be reduced in the future.

Of use in certain circumstances are the stillbirth registrations and, where fertility is of interest as indicating a health effect, the birth registrations may be considered an "endpoint" record. Marriage records could also be used as an intermediate step to get name changes for females. Use of existing marriage and birth files on an historical basis would involve considerable organization and standardization of existing files.

Also potentially useable are the records of hospital admissions and discharges under schemes of universal health insurance (Smith and Newcombe 1975), and the corresponding records of hospital outpatients and visits to physicians' offices. The sheer massiveness of these latter files for historical populations means that considerable expenditure would be involved.

Complementing improvements to the "endpoint" file will be the phase 2 development of the generalized iterative record linkage system. In particular, greater flexibility is needed for generation of a wide variety of reports, and for easily incorporating specific functions tailored to a particular linkage task.

THE NEEDS FOR GENETIC AND FAMILY DATA

For family follow-up of almost any kind, record linkage may be required to establish the family composition. Marriage records, for example, may be used to identify husband-wife pairs; birth registrations may be linked into sibship groupings along with parental marriage records; and the linkage may even be extended for special purposes to identify cousins by arranging the parental birth and marriage records into sibship groupings. Such linkage procedures present no special problems and most of them have been extensively used, notably in British Columbia and later in Hawaii (Newcombe 1969; Trimble 1976; Smith 1980).

For the future one would anticipate that family linkage would be employed where the risks in the workplace are thought to extend into the home, as in the case of asbestos workers, or where the effects of social and economic circumstances, cultural differences, and personal habits such as those associated with diet and smoking might influence similarly the health of all family members (see further examples in Smith 1981).

The possibility that human mutation rates may have been increased by the additional exposures to ionizing radiation and the complex chemicals which accompany industrialization has been repeatedly raised in recent years. Public concern has become a very tangible factor in developing governmental policies and regulations. There is, therefore, a clear need to develop

methods for the identification of populations at possible genetic risk, and a further need to develop consistent and acceptable data bases aimed at determining whether these possible risks, wherever they are presumed to exist, have actually been realized. In particular, it is desirable that possible carcinogenic as well as genetic effects be considered since an increased germinal mutation rate today might be an advanced warning of an increase in malignancy rate 10-20 years hence.

Large studies, (i.e. involving large cohorts), can be carried out at a national level. International co-operation is needed, however, in cases where the numbers are small and pooling of data is required. Two specific examples would have to do with follow-up studies of people who show "early indicator" conditions: (1) to determine among exposed individuals who show, say chromosomal breakage, to see whether they later get cancer, or whether they have reproductive troubles (e.g. stillbirth) or defective children; and (2) to determine among children found by survey to have, say protein variants, whether they later have special health problems or die early.

THE LAW

Statistics Canada does carry responsibility for the confidentiality of the vital and health records which are entrusted to it. There are several pieces of legislation which define what we may or may not do. Most important are the Statistics Act, the recent federal Human Rights Act, Orders-in-Council pertaining to vital statistics, and the various provincial statistics acts or their equivalents. (For further details regarding confidentiality of records see: Medical Research Council 1968; Rowebottom 1979; Krever 1980; Hansen 1981).

CONCLUSION

The approach being taken in Canada, in developing the files and facilities for use in epidemiological studies, is much like the development of the electron microscope for the biologist and others. Here a tool was designed which could serve a multiplicity of users. The facilities should not only meet present needs, but should favour future inquiry into all sorts of questions, without prior knowledge of precisely what these questions will be. The overall strategy for problem solving has been multi-disciplinary. A combination of laboratory, clinical, genetic and epidemiological approaches have contributed in the past, and will be required in the future, to get a clearer understanding of our health problems and their solution.

The recent developments at Statistics Canada will render many more epidemiological studies possible to a degree and a scale of magnitude which have not been seriously contemplated in the past. The technology and systems developed for bringing records together have been well tested. The expansion and enhancement of our current capabilities very much depends on public demand, financial support, and close collaboration and co-operation of a number of federal, provincial and other agencies.

REFERENCES

Ashmore, P., 1979, "The National Dose Register: File Content and Potential Uses," (A paper presented at the NCIC Workshop on 'Computerized Record Linkage in Cancer Epidemiology', Ottawa, 8-10 August 1979).

Dreyer, N.A., et al., 1980, "The Feasibility of Epidemiologic Investigations of the Health Effects of Low-level Ionizing Radiation," Final Report, Health Systems Division Equifax Inc., Reading, MA. 01867. (Report No. NUREG/CR-1728. Available from GPO Sales Program, Division of Technical Information and Document Control, U.S. Nuclear Regulatory Commission, Washington, D.C. 20555).

Hansen, I., 1981, "Report of the Privacy Commissioner on the Use of the Social Insurance Number," Canadian Human Rights Commission, Ottawa, Ontario.

Howe, G.R. and Lindsay, J.P., 1981, "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies," Computers and Biomedical Research (In Press).

Krever, H., 1980, "Report of the Commission of Inquiry into the Confidentiality of Health Information," Ministry of the Attorney General, Province of Ontario, J.C. Thatcher, Queen's Printer, Toronto, Ontario.

Medical Research Council of Canada, 1968, "Health Research Uses of Record Linkage in Canada," Medical Research Council of Canada, Report No. 3, Ottawa, Ontario.

Minister of Supply and Services Canada, 1980, "Mortality Atlas of Canada, Volume 1: Cancer," Catalogue No. H49-6/1-1980, Canadian Government Publication Centre, Hull, Québec. K1A 0S9.

Minister of Supply and Services Canada, 1980a, "Mortality Atlas of Canada, Volume 2: General Mortality," Catalogue No. H49-6/2-1980, Canadian Government Publication Centre, Hull, Québec. K1A 0S9.

Newcombe, H.G., 1969, "The Use of Medical Record Linkage for Population and Genetic Studies," Methods of Information in Medicine, Vol. 8, pp. 7-11.

Newcombe, H.B., 1980, "Design and Future Use of National Dose Registers for Regulatory Control and Epidemiology," Health Physics, Vol. 39, pp. 783-796.

Rowebottom, L.E., 1979, "Statistics Canada: The Legal Basis of Health Record Linkage," (Paper presented at NCIC workshop 'Computerized Record Linkage in Cancer Epidemiology', Ottawa, 8-10 August 1979).

Smith, M.E., 1979, "Automated Medical Follow-up and Delayed Industrial Risks," Proceedings of the First International Conference on Health Effects of Energy Production, N.E. Gentner and P. Unrau, eds., Report No. AECL 6958, Chalk River Nuclear Laboratories, Chalk River, Ontario, pp. 125-132.

Smith, M.E., 1980, "The Present State of Automated Follow-up in Canada, Part 1: Methodology and Files," Journal of Clinical Computing, Vol. 9, pp. 1-18.

Smith, M.E., 1981, "Value of Record Linkage Studies in Identifying Populations at Genetic Risk and Relating Risk to Exposures," Proceedings of an International Symposium on Chemical Mutagenesis, Human Population Monitoring, and Genetic Risk Assessment, K.C. Bora, ed., Elsevier/ North Holland Biomedical Press, B.V. Amsterdam (In press).

Smith, M.E., 1981a, "Long-term Medical Follow-up in Canada," A paper presented at a conference on the Quantification of Occupational Cancer, Banbury Center, 29 March - 2 April 1981. Proceedings of the conference are to be published in Banbury Report No. 9 - Quantification of Occupational Cancer, Banbury Center, Cold Spring Harbor Laboratory, P.O. Box 534, Cold Spring Harbor, New York. 11724.

Smith, M.E. and Newcombe, H.B., 1975, "Methods for Computer Linkage of Hospital Admission - Separation Records into Cumulative Health Histories," Methods of Information in Medicine, Vol. 14, pp. 118-125.

Smith, M.E. and Newcombe, H.B., 1981, "Use of the Canadian Mortality Data Base for Epidemiological Follow-up," Canadian Journal of Public Health (In press).

Smith, M.E. and Silins, J., 1981, "The Generalized Iterative Record Linkage System," A paper presented at the American Statistical Association Conference, Detroit, August 9-13, 1981.

Smith, M.E., Silins, J. and Lindsay, J.P., 1980, "The Present State of Automated Follow-up in Canada, Part II: The Products," Journal of Clinical Computing, Vol. 9, pp. 19-30.

Trimble, B.K., 1976, "Record Linkage and Genetic Registers," Registers for the Detection and Prevention of Genetic Disease, A.E.H. Emery and J.R. Miller, eds., Symposia Specialists, 1460 NE 129 Street, Miami, Florida. 33161.

POOR

QUALITY

ORIGINAL